APPLYING STATISTICAL GRAPHICS TO MULTIVARIATE DATA.

Jan. 1986

Steven J. Schwager
Biometrics Unit
337 Warren Hall
Cornell University
Ithaca, NY 14853-0401

ABSTRACT

Graphical techniques for displaying, examining, and analyzing multivariable observations are discussed. Graphical methods that reveal important features of data serve to complement and illuminate formal statistical inferences. Recently developed graphical displays having practical value for applied work with high-dimensional data are emphasized. Star plots, faces, and trees are examples of such methods. The strengths and weaknesses of these and other techniques for dealing with data from applied situations will be treated and compared.

INTRODUCTION

A major goal of graphical methods for statistical data is to make large amounts of numerical information comprehensible. While other statistical methods are also directed at this goal, graphical techniques seek to achieve it by displaying the essential features of the data in visual form. The fundamental problem of statistical graphics is to discover the display or displays that reveal the most about the data's important features.

Graphical methods can be used to advantage in any of three broad statistical activities. The first of these is exploration, the examination of data by judicious display and description, in an attempt to determine the overall structure of the data. The second is analysis, including summarization, the distillation of a few statistics or displays that adequately reflect the data's main characteristics; exposure, the revelation of any unexpected features present in the data; statistical modeling, the selection of a specific mathematical pattern associated with the data; model testing, the examination of whether the data are in fact compatible with the chosen model; and other similar operations. The third activity is the communication of results to a suitable audience, for whom the crucial features of the data should be displayed with clarity and simplicity. These activities often overlap, and the statistician often iterates back and forth among them until desirable results are achieved. Graphical methods can make an important contribution in each of these areas.

The problem to be considered in this paper is the use of graphical techniques to represent multivariate observations. Data of this kind consist of n observations, on each of which a set of p variables has been measured. The observations are also called observational units or cases; they may be individuals, geographic entities like cities or countries, instants or periods of time, automobile models, kinds of food, or any collection of comparable items. The variables, also called dimensions or components, are the aspects of the observations that have been measured. For example, if the observations are automobile models, the variables might be gasoline mileage, weight, number of cylinders, engine displacement, and so on. If the

observations are cities, the variables might be population, per capita income, number of dwelling units available for occupancy, rates of occurrence of various kinds of crime, and so on. Graphical techniques offer an approach to the investigation of many variables simultaneously. The values of the p variables for a single observation may be represented visually as a graphical display, and the n displays, one for each observation, may be examined together for evidence of the underlying structure of the data.

The graphical techniques considered here are tools of what Gnanadesikan (1973) calls "informal inference." These are techniques by which data can be explored and insights about the data's structure can be uncovered and displayed. Gnanadesikan notes the complementary nature of these informal techniques and the methods of formal statistical inference: graphical representations of multivariate data should be used "in conjunction with, and as aids for," numerical computations and analyses rather than as substitutes for these. The availability of increasingly powerful and flexible computer graphics improves the quality of our tools, but does not eliminate the need for other tools of different kinds.

Cox (1978) gives a set of guidelines for statistical graphics. Several of these are especially important to keep in mind while addressing the problem of displaying n observations, each consisting of p variables. The labeling of variables should be clear, including the variable names and units of measurement. Legends should make diagrams as nearly self-explanatory as possible. Comparison of related diagrams should be easy. Interpretation should not be prejudiced by the techniques used in presentation.

Recent general treatments of statistical graphics include books by Bertin (1983), Chambers, Cleveland, Kleiner, and Tukey (1983), Schmid (1983), and Tufte (1983), and articles by Gentleman (1983) and Snee and Pfeifer (1983). If anyone should doubt that graphical methods can have an important place in statistical analyses, the classic example of Anscombe (1973) demonstrates forcefully that very different-looking data sets can produce identical regression analyses and summary statistics. Without visual inspection, the procedures of formal statistical inference can fail to reveal extreme differences among data sets.

The discussion here will exclude methods that involve dimension reduction, in which the p variables measured on each observation are reduced to a smaller collection of p' variables, where p' is less than p. It will also exclude methods that rely on color or intensity of shading, and those that rely on the use of motion. Each of these approaches introduces additional problems to the development of graphical displays. Dimension reduction requires methods for determining an appropriate value of p' (less than p) and effectively transforming the original p variables into p' new variables that retain nearly all of the information present. The use of color introduces two problems mentioned by Bertin (1983, p. 91): the plight of people who have some form of color blindness, which is "more frequent than is generally believed," and the practical difficulties of reproducing photocopies in color. The latter problem is shared by methods that depend on motion to display multivariate data.

The objective here is to treat suitable statistical graphics for displaying all p variables for all n observations. The display may be intended for purposes of exploration, analysis, communication, or some combination of these aims. The value of p may change as work progresses, for it may become clear that some of the p variables are superfluous and should be eliminated from consideration, or it may become attractive to add new variables not originally anticipated. However, at any given stage of work, the display of all p variables for all n observations as informatively as possible will be what we wish to achieve.

The range of n envisioned here is from a few observations to a few hundred; the range of p is from 2 to about 20. The cost in effort, both computational and mental, increases with both n and p, so the graphical methods to be treated cannot be considered equally applicable to all combinations of n and p. Unfortunately, and inevitably, the ability of graphical methods to reveal the underlying structure of the data depends on the complexity of this structure, which is unknown when examination of the data commences (and sometimes when it concludes). Consequently, the values of n and p and the complexity of the relationships among variables and observations all influence how helpful a given graphical technique will be and how much effort will be required to utilize it. This assessment should not be construed as unduly pessimistic; after all, even when much effort is required and relatively little insight is gained from a graphical display, nongraphical techniques are likely to produce an even smaller return for an equal or greater expenditure of effort.

For convenience, the term "analysis" in phrases like data analysis and graphical analysis will be used from this point on to mean the combined operations of exploration, analysis, and communication described earlier in this section.

PRELIMINARY STEPS FOR MULTIVARIATE GRAPHICS

Any data analysis that omits or mishandles the initial steps is unlikely to have an appropriately happy ending. A critical matter that requires attention at the outset is defining the goal of the analysis. Many possible goals can be described, for example: (1) to "see" in a loose, informal way the structure of the data set; (2) to partition the n observations into several nonoverlapping sets called clusters, each of which is relatively homogeneous; (3) to investigate the relationship between the p given variables and an externally specified variable, such as the geographic location of each of the n observations. Numerous other goals can be proposed, some exploratory in nature, like (1), others concerned with analytical issues, like (2), and still others concerned with presentation, like (3). The explicit definition of a goal is vitally important because a graphical or statistical technique that is appropriate and productive for achieving a particular goal may be inappropriate and unproductive for achieving a different goal. For this reason, specifying the goal must precede meaningful analysis of the data.

Screening the data for anomalies and unexpected patterns is another critical matter. A single outlier, or aberrant observation, can have enormous effects on both graphical and inferential analyses; more pervasive anomalies can be even more destructive in their consequences. Three methods of preliminary screening of observations will be briefly described. First, univariate histograms of the n observed values of each of the p variables will reveal outliers, clusters, asymmetry, and many other unusual phenomena if they are present. (Note that outliers may be incorrect values caused by measurement or recording error, such as keying mistakes in data entry, or correct values that constitute unusual outcomes, or of unknown cause; the appropriate procedure for handling outliers depends on the circumstances, as Barnett and Lewis (1984) discuss. Similar considerations apply to other anomalies.) Second, patterns involving two variables can be seen from bivariate scatterplots of pairs of variables, each plot containing n points corresponding to the observations. Chambers, Cleveland, Kleiner, and Tukey (1983, Ch. 5) refer to the set of all such bivariate plots as a generalized draftsman's display. They also discuss enhanced scatterplots: a bivariate plot can be labeled with a symbol representing the value of a third variable; the joint behavior of four variables can be shown using a plot called a multiwindow display. These plots will reveal many unusual and troublesome patterns when they occur, but two cautions must be sounded. Inspecting the large number of plots (p variables give $p(p+1)/2$ bivariate plots) can lead to confusion. Even worse, the structure of the full set of p variables, viewing the

observations as n points in p-dimensional space, may not be accurately portrayed by plots of two, three, or four variables. Everitt (1978) cites data given by Cattell and Coulter (1966) and Nathenson (1971) demonstrating this. The third method of preliminary data screening is to use graphical displays for multivariate data, which will be described in the next section, in an exploratory examination of the data set's structure. If an observation is an outlier, its graphical display will differ greatly in appearance from the rest.

Transformations of the data, especially those suggested by the data, will often make the structure more visible, which improves our chances of detecting and understanding it. The selection of beneficial transformations, called "first aid" by Tukey and Mosteller (1977), is as much an art as a science. Logarithms, sign changes, and power transformations are helpful in many cases. An especially useful transformation is changing the signs of certain variables to make the relationships among strongly associated variables increasing instead of decreasing.

GRAPHICAL DISPLAYS FOR MULTIVARIATE DATA

A graphical representation of n observations, each consisting of measurements on p variables, will contain n display graphics, which we will call symbols, one for each observation. The configurations of these symbols can be compared visually, leading to a graphical analysis of the multivariate data from which the symbols came. We now present eleven methods of transforming a p-dimensional observation into a display graphic or symbol. Each method is actually a family of rules rather than a single rule for producing a symbol, because choices required by the method are made by the user, either explicitly or implicitly. These choices will be mentioned as the methods are discussed. Examples of nine of the eleven methods can be found in one or more of Chambers, Cleveland, Kleiner, and Tukey (1983), Gnanadesikan (1977), Kleiner and Hartigan (1981), and Tukey and Tukey (1981). The remaining two methods are constellations and asymmetric faces.

Glyphs or Metroglyphs. (Anderson, 1957) Each variable is represented as a ray extending out from a circle of fixed size, the length of the ray corresponding to the value of the variable. The assignment of variables to rays on the circle is, in general, arbitrary. The glyph's appearance depends on this assignment, especially as the number of variables increases, and also on the rules that relate variable values to ray lengths. We will assume that ray length is a linear function of variable value, as is usually done.

Profiles. (Bertin, 1967) The value of each variable is transformed to a height above a horizontal line. The observation is then represented either by connecting these p heights, which are equally spaced horizontally, to form a polygonal line or by forming a histogram from the p bars drawn at the heights obtained from the variables. The profile's appearance depends on the order of the variables, which is, in general, arbitrary. Comparisons among observations are difficult, particularly when there are many variables.

Stars, Polygons, or Sun Ray Plots. (Goldwyn, Friedman, and Siegel, 1971) Each variable is represented as a point on a ray emanating from the center of a circle. These points on the p equally spaced rays are connected, forming a star or polygon. This symbol is a circular version of a profile. Its appearance depends on the order of the variables, which is arbitrary. Comparisons among observations are difficult, particularly when there are many variables. When adjacent variables are not strongly related and there are many variables, stars are jagged and hard to interpret.

Weathervane Plots. (Bruntz, Cleveland, Kleiner, and Warner, 1974) Each symbol consists of a circle with a ray extending from its center. The diameter of the circle is proportional to observed daily maximum temperature, the direction of the ray gives the day's average wind direction, and the length of the ray is inversely proportional to average daily wind speed. Each observation is a period of one day. More elaborate versions of the weathervane plot symbol could be developed for more than three variables. The features of the graphic display correspond to the variables in a natural way here, but in general this will not be the case. In addition, complications would result from the addition of several more variables, which would require a more detailed symbol.

Fourier Plots. (Andrews, 1972) Each observation is represented by the plot of a function on an interval. The function is a linear combination of the trigonometric quantities $\sin(t)$, $\cos(t)$, $\sin(2t)$, $\cos(2t)$,... with coefficients given by the values of the p variables. Thus each observation is represented as a curve composed of sines and cosines multiplied by the p values of the variables and summed. It is usually informative to plot all n of these curves on a single diagram. Variations of this curve have been suggested (e.g., see Gnanadesikan, 1977, pp. 207-209). All of them depend on the order of the variables, which is arbitrary.

Faces. (Chernoff, 1973) Each variable corresponds to one or more features of a face: eye size, eye slant, eyebrow curvature, lower hair line location, mouth size, and so on. The p variable values constituting an observation determine all facial characteristics, producing a face that represents the observation. Comparisons among observations can be relatively easy because viewers are able to make fairly good judgments about similarity and difference based on facial features. However, the performance of these faces depends on the correspondence between variables and facial features, especially when there are many variables.

Boxes. (Hartigan, 1975) Each variable is represented as the length of a rectangular box in one of the three dimensions. If there are more than three variables, one or more of the three dimensions will be divided into several shorter segments. Each observation appears as a box, with wrapping strings if p exceeds 3. These strings divide the edges of the box into segments whose lengths correspond to the values of the variables. The box's appearance depends on the assignment of the variables. So does the difficulty of making comparisons among observations, which increases with the number of variables.

Constellations. (Wakimoto and Taguri, 1978) The observations are transformed, component by component, to make every variable observed for every observation lie between o and π. (A linear transformation of the n observed values for each variable is a straightforward way to accomplish this.) Each variable of a particular observation is then replaced by its sine and cosine; a weighted sum of these p sines is computed, as well as a weighted sum of these p cosines with the same weights. Wakimoto and Taguri propose a bivariate scatterplot of this pair of weighted sums for each observation, reducing each p-dimensional observation to a bivariate quantity. They also consider plotting for a particular observation the weighted sum of the first $1,2,\ldots,p$ sines against the weighted sum of the first $1,2,\ldots,p$ cosines, respectively, and connecting the origin and these p points to obtain a path. Each observation can be represented by such a path. The path depends on the order of the p variables, the initial transformation used on the data, and the weights used in the sums of sines and cosines.

Trees. (Kleiner and Hartigan, 1981) It will be assumed that all variables take comparable values; standardize the variables if this is necessary to achieve comparability. Then perform a hierarchical

clustering of the p variables, to group them according to similarity. The tree diagram resulting from this clustering is used to construct a template or prototype tree, in which the thickness of each branch, the angles between branches, and similar quantities have been determined. Finally, construct a tree for each observation, in which: the correspondence of variables to branch locations, the branch thickness, the angles between branches, and so on conform exactly to the prototype tree; and the branch lengths of the tree representing a given observation are computed from the values of the variables for that observation. An observation with low values of some variables will have short branches in the section of its tree corresponding to those variables; high values of some variables will result in long branches in the corresponding section of the tree for this observation. The hierarchical clustering of variables eliminates the dependence of the n displays (one for each observation) on the order of the variables. Comparisons among observations are relatively easy. Comparisons of variables within the same tree, however, are difficult.

Castles. (Kleiner and Hartigan, 1981) The castle, which is a combination tree and profile, makes it easy to compare different variables from a single observation, and to compare the values of a particular variable in several observations. As with trees, assume that all variables are comparable, perhaps after standardizing. Perform a hierarchical clustering of the variables to order them, then construct trees having angles of zero between all pairs of branches and branch thickness proportional to the number of variables included in the branch. With a suitable rule for determining branch lengths, the castles become identical to profiles, except that the order of the variables has changed because of the clustering. This makes the castles more informative than profiles and removes the dependence of the n displays on the order of the variables.

Asymmetric Faces. (Flury and Riedwyl, 1981) These symbols are a variation of Chernoff's faces in which there are 18 parameters for each side of the face, so the left and right halves together have 36 parameters, each corresponding to a facial feature. When there are less variables than parameters, some variables can control several parameters, thereby determining several facial features, just as can occur with symmetric faces. Of course, selected parameters can be held fixed for all n observations if one chooses. Associating a pair of positively correlated variables with the same feature on the left side and on the right side will avoid severe asymmetries. The behavior of asymmetric faces is quite similar to that of symmetric faces. Asymmetric faces can accommodate more variables, up to 36 per observation; however, severe asymmetries in the n graphical displays can be distracting to the viewer, so care must be taken in establishing the correspondence between the p variables and facial features.

ENHANCEMENTS OF MULTIVARIATE GRAPHICAL DISPLAYS

The graphical methods just described can be improved in three ways we now consider.

Combining Graphical Symbols and Rectangular Coordinates

We have viewed glyphs, stars, faces, trees, and other symbols as objects without any fixed position or location. The n observations, each consisting of p variables, are represented by n symbols, each of which depicts the values of the p variables for one observation. This leaves us free to move the symbols about as we like. For example, we can sort the symbols into some natural order or divide them into several relatively homogeneous clusters, using general visual impressions of the symbols, a quantitative measure of the symbols' behavior, or any other means. We can issue commands at a graphics terminal that will physically move the symbols to wherever we want them.

We may instead choose to derive a position in the xy-plane from two of the variables and to plot at that position a symbol representing the values of the other p-2 variables. Doing this for each of the n observations will give a scatterplot in which the plotted objects are graphical symbols rather than points. Glyphs, stars, and weathervanes are the symbols usually seen in plots combining planar position with graphical displays. However, any of the symbols described in the preceding section could appear in such a plot. Tukey and Tukey refer to these and similar symbols, in which several variables are represented by a single graphical symbol for each observation in the data set, as "individual-value compound characters" (Barnett, 1981, pp. 256-257). The question of which two variables to associate with position in the plane must be added to the usual question of which graphical symbol and assignment of variables to its features will be most informative.

A special case of graphical display with planar position is the geographic display map. In this application a graphical representation of p-dimensional data is superimposed on a geographical area: state by state within the U. S., census tract by census tract within a city, etc. The graphical symbol used can be stars, faces, or even a figure we have not mentioned. Schmid (1983, pp. 120, 144) advocates using pie charts, with the size of the pie proportional to the district's total count or amount and the area of each slice proportional to a category count or amount (where the p variables are numbers of residents of p kinds, amounts of money spent for p purposes, or any p comparable quantities). Conversely, Tufte (1983, p. 178) says, "Given their low data-density and failure to order numbers along a visual dimension, pie charts should never be used." Bertin (1983, pp. 118-119, 139-151) agrees.

Permutation of Variables

The assignment of variables to the specific features of a graphical symbol has great influence on the effectiveness of the resulting display. This was seen earlier: the appearance of many of the symbols was heavily dependent on the order in which the variables are assigned to the features of the symbol. Changing this assignment can improve the information content of a graphical display substantially. Clustering of variables, as in the derivation of trees and castles, moves similar variables close to each other. A preliminary clustering of variables should be expected to improve the performance of graphical symbols like glyphs and stars, as Jacob (1981) has noted. Hierarchical clustering allows a wide choice of distance metric (Euclidean, city block, sup norm, Mahalanobis, and other metrics) and amalgamation rule (single, average, and complete linkage, and others); some combinations are likely to perform much better than others in making graphical displays more informative. It may be difficult to know in advance which will do well. When the symbol is faces, some assignments of variables to features will give much better results than others. Methods for determining these assignments would be of great interest.

Implementation

Graphical methods for multivariate data will benefit from wider availability, greater ease of use, and greater flexibility.

There is a need for issues of perception to be addressed more fully. The format and style of graphical displays should be as helpful to the viewer as possible. For example, faces appear in two styles, "quasirealistic" faces and "cartoon" faces. Is one of these styles much easier to examine and more informative? If so, which style is it? Questions in a similar spirit can be raised about other graphical methods for multivariate data.

PERFORMANCE OF GRAPHICAL METHODS FOR MULTIVARIATE DATA

The question of how well graphical methods for multivariate data behave has received some attention. Chernoff and Rizvi (1982) examined how accurately subjects were able to divide a set of 36 faces into two groups of approximately equal size with internal similarity within each group. They concluded that randomly permuting the assignment of variables to facial features leads to fluctuations in error rate with a factor of about 25%. Everitt (1978, pp. 90-94) illustrated that different assignment of features can produce a set of faces that makes it more difficult to extract information and to see the structure of the data set. Freni-Titulaer and Louv (1984) found that trees outperformed castles, histograms, and clustered histograms when subjects were asked to divide a set of 16 observations into two internally similar groups of eight. This is a start, but much more remains to be done.

Sensitivity to the order of the variables, which controls the assignment of variables to the display features, is a major concern. Glyphs, profiles, and stars seem likely to be relatively sensitive to order.

The ability of graphical displays of multivariate data to inform us is influenced by many factors, including the values of n and p, the (unknown) structure underlying the data, and the assignment of variables to display features. Initial results suggest that faces and trees are the most informative displays, but definitive answers are not yet in sight.

REFERENCES

Anderson, E. (1957). A semigraphical method for the analysis of complex problems. Proc. Nat. Acad. Sci. USA, 13, 923-927. Reprinted in Technometrics (1960) 2, 387-391.

Andrews, D. F. (1972). Plots of high-dimensional data. Biometrics 28, 125-136.

Anscombe, F. J. (1973). Graphs in statistical analysis. The American Statistician 27, 17-21.

Barnett, V. (ed) (1981). Interpreting Multivariate Data. Wiley, Chichester, U. K.

Barnett, V. and Lewis, T. (1984). Outliers in Statistical Data, 2nd edition. Wiley, New York.

Bertin, J. (1967). Semiologie Graphique. Gauthier-Villars, Paris.

Bertin, J. (1983). Semiology of Graphics. University of Wisconsin Press, Madison.

Bruntz, S. M., Cleveland, W. S., Kleiner, B., and Warner, J. L. (1974). The dependence of ambient ozone on solar radiation, wind, temperature, and mixing height. Proc. Symp. Atmos. Diffus. Air Pollution, American Meteorological Society, 125-128.

Cattell, R. B. and Coulter, M. A. (1966). Principles of behavioural Taxonomy and the mathematical basis of the taxonome computer program. British Journal of Mathematical and Statistical Psychology 19, 237-269.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P.A. (1983). Graphical Methods for Data Analysis. Wadsworth, Belmont, California.

Chernoff, H. (1973). The use of faces to represent points in k-Dimensional space graphically. Journal of the American Statistical Association 68, 361-368.

Chernoff, H. and Rizvi, M. H. (1975). Effect on classification error of random permutations of features in representing multivariate

data by faces. Journal of the American Statistical Association 70, 548-554.

Cox, D. R. (1978). Some remarks on the role in statistics of graphical methods. Journal of the Royal Statistical Society, Series C 27, 4-9.

Everitt, B. S. (1978). Graphical Techniques for Multivariate Data. North-Holland, New York.

Flury, B. and Riedwyl, H. (1981). Graphical representation of multivariate data by means of asymmetrical faces. Journal of the American Statistical Association 76, 757-765.

Gentleman, J. E. (1983). Graphical representation, computer aided. In Encyclopedia of Statistical Sciences, Vol. 3, Kotz, S. and Johnson, N. L. (eds), Wiley, New York.

Gnanadesikan, R. (1973). Graphical methods for informal inference in multivariate data analysis. Bull. Int. Stat. Inst., Proc. 39th Sess. ISI at Vienna 45, Book 4, 195-206.

Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. Wiley, New York.

Goldwyn, R. M., Friedman, H. P., and Siegel, T. H. (1971). Iteration and interaction in computer data bank analysis; case study in physiological classification and assessment of the critically ill. Computers in Biomedical Research 4, 607-622.

Hartigan, J. A. (1975). Printer graphics for clustering. Journal of Statistical Computation and Simulation 4, 187-213.

Jacob, R. J. K. (1981). Comment on "Representing points in many dimensions by trees and castles" by Kleiner and Hartigan. Journal of the American Statistical Association 76, 270-272.

Kleiner, B. and Hartigan, J. A. (1981). Representing points in many dimensions by trees and castles. Journal of the American Statistical Association 76, 260-269.

Nathanson, J. A. (1971). Applications of multivariate analysis in astronomy. Applied Statistics 20, 239-249.

Schmid, C. F. (1983). Statistical Graphics. Wiley, New York.

Snee, R. D. and Pfeifer, C. G. (1983). Graphical representation of data. In Encyclopedia of Statistical Sciences, Vol. 3, Kotz, S. and Johnson, N. L. (eds), Wiley, New York.

Tufte, E. R. (1983). The Visual Display of Quantitative Information. Graphics Press, Cheshire, Connecticut.

Tukey, J. W. and Mosteller, F. (1977). Data Analysis and Regression. Addison-Wesley, Reading, Massachusetts.

Tukey, P. A. and Tukey, J. W. (1981). Summarization; smoothing; supplemented views. In Interpreting Multivariate Data, Barnett, V. (ed), Wiley, Chichester, U. K.

Wakimoto, K. and Taguri, M. (1978). Constellation graphical method for representing multi-dimensional data. Annals of the Institute of Statistical Mathematics 30, 97-104.